Caching in Wireless Networks

Urs Niesen, Devavrat Shah, Gregory Wornell

Abstract

We consider the problem of delivering content cached in a wireless network of n nodes randomly located on a square of area n. In the most general form, this can be analyzed by considering the $2^n \times n$-dimensional caching capacity region of the wireless network. We provide an inner bound on this caching capacity region, and, in the high path-loss regime, a matching (in the scaling sense) outer bound. For large path-loss exponent, this provides an information-theoretic scaling characterization of the entire caching capacity region. Moreover, the proposed communication scheme achieving the inner bound shows that the problem of cache selection and channel coding can be solved separately without loss of order-optimality.

I. Introduction

With the continued large-scale deployment of infrastructure, wireless networking continues to be an area of active research. In this context, unicast and multicast traffic has been widely studied. The influence of caches on the network performance, on the other hand, has received considerably less attention. Nevertheless, the ability to replicate data at several places in the network is likely to significantly increase supportable data rates. In this paper, we consider the problem of characterizing achievable rates with caching in large wireless networks.

In its most general form, this problem can be formulated as follows. Consider a wireless network with n nodes, and assume a node w in the network requests a message available at the set of caches U (a subset of the n nodes) at a certain rate $\lambda^{\mathsf{CA}}_{U,w}$. The collection of all $\{\lambda^{\mathsf{CA}}_{U,w}\}_{U,w}$ can be represented as a caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$. The question is then to characterize the set of achievable caching traffic matrices $\Lambda^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$. We answer this question by providing an approximate (i.e., scaling) characterization of this caching capacity region $\Lambda^{\mathsf{CA}}(n)$

This work was supported, in part, by DARPA under Grant No. 18870740-37362-C (ITMANET), and by NSF under Grant No. CCF-0635191.

for large wireless networks (i.e., as $n \to \infty$) under random node placement and assuming large path-loss exponent. Our treatment is information-theoretic, i.e., we do not make any assumptions on the communication protocol used.

While $\Lambda^{\mathsf{CA}}(n)$ is a high-dimensional object (namely $2^n \times n$-dimensional), we show that feasibility of a traffic matrix λ^{CA} can be efficiently evaluated. We also provide an explicit communication scheme achieving (in the scaling sense) the entire caching capacity region $\Lambda^{\mathsf{CA}}(n)$.

A. Related Work

Several aspects of caching in wireless networks have been investigated in prior work. In the computer science literature, the wireless network is usually modeled as a graph induced by the geometry of the node placement. This is tantamount to making a protocol model assumption (as proposed in [1]) about the communication scheme used. The quantity of interest involves the distance from each node to the closest cache that holds the requested message. The problem of optimal cache location for multicasting from a single source has been investigated in [2], [3]. Optimal caching densities under uniform random demand have been considered in [4], [5]. Several cache replacement strategies are proposed, for example, in [6].

To the best of our knowledge, caching has not been directly considered in the information theory literature. However, it can be seen that the problem of optimally transmitting messages held at several caches to a destination is a special case of communicating correlated sources over a noisy network. Indeed, we can consider that each cache has an identical message to send to the same destination. This more general problem of transmitting correlated sources has received considerable attention. Unlike the situation with point-to-point communication, for network communication problems source-channel separation does not hold in general [7]. Hence, the problem of source and channel coding have to be considered jointly. While for some special cases optimal communication strategies for transmitting correlated sources over a noisy network are known (for example, a single destination node requesting all the sources observed in the network with independent network links [8], [9]), the general problem is unsolved.

Finally, a special case of the caching problem considered here, in which each destination has only a single cache (i.e., standard unicast traffic), has been widely studied and is by now well understood. See, for example, [1], [10]–[20].

B. Our Contribution

We consider the general caching problem from an information-theoretic point of view. Compared to the prior work mentioned in the last section, there are several key differences. First, we do not make a protocol channel model assumption, and instead allow the use of arbitrary communication protocols over the wireless network. Second, we allow for general traffic demands, i.e., arbitrary number of caches, and arbitrary demands from each destination. Third, we do not impose that each destination requests the desired message from only the closest cache, nor do we impose that the entire message has to be requested from the same cache. Rather we allow parts of the same message to be requested from distinct caches.

We present an achievable communication scheme for the caching problem, yielding an inner bound on the caching capacity region. For large values of path-loss exponent, we provide a matching (in the scaling sense) outer bound, proving the optimality (again in the scaling sense) of our proposed scheme. Together, this provides a scaling description of the entire caching capacity region of the wireless network in the large path-loss regime. The proposed communication scheme solves the problem of optimal cache selection and channel coding separately, showing that such a separation is order-optimal.

C. Organization

The remainder of the paper is organized as follows. In Section II, we introduce the channel model as well as notation. In Section III, we present the main results of the paper. Section IV contains proofs, and Section V concluding remarks.

II. Network Model and Notation

Consider the square

$$A(n) \triangleq [0, \sqrt{n}]^2$$

of area n, and let $V(n) \subset A(n)$ be a set of $|V(n)| = n$ nodes on $A(n)$. We assume the following channel model. The (sampled) received signal at node v and time t is

$$y_v[t] = \sum_{u \in V(n) \setminus \{v\}} h_{u,v}[t] x_u[t] + z_v[t]$$

for all $v \in V(n), t \in \mathbb{N}$, and where $\{x_u[t]\}_u$ are the (sampled) signals sent by the nodes in $V(n)$ at time t. Here $\{z_v[t]\}_{v,t}$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian random variables with mean 0 and variance 1, and

$$h_{u,v}[t] = r_{u,v}^{-\alpha/2} \exp(\sqrt{-1}\theta_{u,v}[t]),$$

for *path-loss exponent* $\alpha > 2$, and where $r_{u,v}$ is the Euclidean distance between u and v. The phase terms $\{\theta_{u,v}[t]\}_{u,v}$ is assumed to be i.i.d. with uniform distribution on $[0, 2\pi)$.[1] We either assume that $\{\theta_{u,v}[t]\}_t$ is stationary and ergodic as a function of t which is called *fast fading* in the following, or we assume $\{\theta_{u,v}[t]\}_t$ is constant as a function of t, which is called *slow fading* in the following. In either case, we assume full channel state information (CSI) is available at all nodes, i.e., each node knows all $\{h_{u,v}[t]\}_{u,v}$ at time t.[2] We also impose an average power constraint of 1 on the signal $\{x_u[t]\}_t$ for every node $u \in V(n)$.

Partition $A(n)$ into 4^ℓ subsquares $\{A_{\ell,i}(n)\}_{i=1}^{4^\ell}$ of sidelength $2^{-\ell}\sqrt{n}$, and let $V_{\ell,i}(n)$ be the nodes in $A_{\ell,i}(n)$. The integer parameter ℓ varies between 0 and $L(n)$ defined as[3]

$$L(n) \triangleq \frac{1}{2}\log(n)\big(1 - \log^{-1/2}(n)\big).$$

The partitions at various levels ℓ form a dyadic decomposition of $A(n)$ as illustrated in Figure 1.

A *caching traffic matrix* is an element $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$. Consider $U \subset V(n)$ and $w \in V(n)$. Assume a message that is requested at destination node w is available at all of the caches U. $\lambda_{U,w}^{\mathsf{CA}}$ denotes then the rate at which node w wants to obtain the message from the caches U. Note that we do not impose that any particular cache $u \in U$ provides w with the desired message, rather multiple of the nodes in U could provide parts of the message. Note also that $\lambda_{U,w}$ and

[1]It is worth pointing out that recent results [20] suggest that, under certain assumptions on the location of scattering elements, for $\alpha \in (2, 3)$ and very large values of n, the channel model used here (in particular, the i.i.d. assumption of the phase terms, might yield results that are too optimistic. However, in [21] the same authors show that, under different assumptions on the scatterers, the channel model used here is still valid also for $\alpha \in (2, 3)$ and very large values of n. This indicates that the issue of proper channel modeling in the low path-loss regime for very large networks is somewhat delicate and requires further investigation.

[2]We make the full CSI assumption in all the converse results in this paper. Achievability can be shown to hold under weaker assumptions on the availability of CSI. In particular, for $\alpha \geq 3$, no CSI is necessary, and for $\alpha \in (2, 3)$, a 2 bit quantization of the channel state $\{\theta_{u,v}[t]\}_{u,v}$ available at all nodes at time t is sufficient.

[3]Throughout $\log$ and $\ln$ represent the logarithms with respect to base 2 and e, respectively.

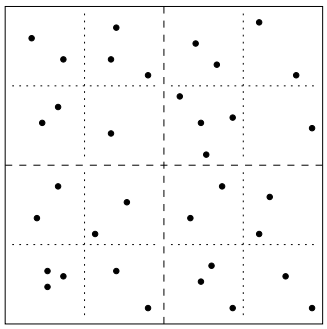

Fig. 1. Subsquare $\{A_{i,\ell}\}$ with $0 \leq \ell \leq 2$, i.e., with $L(n) = 2$. The subsquare at level $\ell = 0$ is the area $A(n)$ itself. The subsquares at level $\ell = 1$ are indicated by dashed lines, the subsquares at level $\ell = 2$ by dotted lines. Assume for the sake of example that the subsquares are numbered from left to right and then from bottom to top (the precise order of numbering is immaterial). Then $V_{0,1}(n)$ are all the nodes $V(n)$, $V_{1,1}(n)$ are the nine nodes in the lower left corner (delineated by dashed lines), and $V_{2,1}(n)$ are the three nodes in the lower left corner (delineated by dotted lines).

$\lambda_{\widetilde{U},w}$ could both be strictly positive for $U \neq \widetilde{U}$, i.e., the same destination could request more than one message from different collection of caches. We assume that messages for different (U, w) pairs are independent. The *caching capacity region* $\Lambda^{\mathsf{CA}}(n)$ of the wireless network $V(n)$ is the set of all achievable caching traffic matrices $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$.

Example 1. Consider $V(n) = \{v_i\}_{i=1}^4$ with $n = 4$. Assume that v_1 requests a message $m_{\{v_3,v_4\},v_1}$ available at the caches v_3, and v_4 at rate 1 bit per channel use, and an independent message $m_{\{v_3\},v_1}$ available only at v_3 at a rate of 2 bits per channel use. Node v_2 requests a message $m_{\{v_3,v_4\},v_2}$ available at the caches v_3 and v_4 at a rate of 4 bits per channel use. The messages $m_{\{v_3,v_4\},v_1}$, $m_{\{v_3\},v_1}$, and $m_{\{v_3,v_4\},v_2}$ are assumed to be independent. This traffic pattern can be described by a caching traffic matrix $\lambda \in \mathbb{R}_+^{16\times 4}$ with $\lambda_{\{v_3,v_4\},v_1} = 1$, $\lambda_{\{v_3\},v_1} = 2$, $\lambda_{\{v_3,v_4\},v_2} = 4$, and $\lambda_{U,w} = 0$ otherwise. Note that in this example node v_1 is destination for two (independent) caching messages, and node v_3 and v_4 serve as caches for more than one message (but these messages are again assumed independent). $\Diamond$

To simplify notation, we will assume when necessary that large reals are integers and omit $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ operators. For the same reason, we suppress dependence on n within proofs whenever this dependence is clear from the context. We use bold font to denote matrices whenever the matrix structure is of importance. We use the $\dagger$ symbol to denote the complex conjugate of a

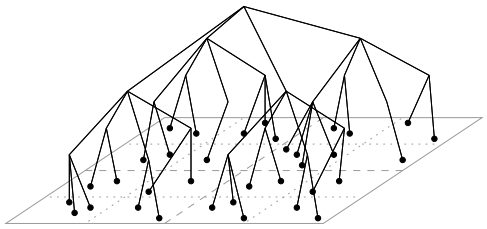

Fig. 2. Construction of the tree graph G. We consider the same nodes as in Figure 1 with $L(n) = 2$. The leaves of G are the nodes $V(n)$ of the wireless network. They are always at level $\ell = L(n) + 1$ (i.e., 3 in this example). At level $0 \leq \ell \leq L(n)$ in G, there are 4^ℓ nodes. The tree structure is induced by the decomposition of $V(n)$ into subsquares $\{V_{\ell,i}(n)\}_{\ell,i}$, delineated by dashed and dotted lines. Level 0 contains the root node of G.

matrix.

III. Main Results

We now present the main results of this paper. In Section III-A, we provide an inner and, for large path-loss exponents $\alpha > 6$, a matching (in the scaling sense) outer bound on the capacity region $\Lambda^{\mathsf{CA}}(n)$. In Section III-B, we discuss computational aspects. In Section III-C, we introduce the communication scheme achieving the inner bound on $\Lambda^{\mathsf{CA}}(n)$. We analyze several example scenarios in Section III-D.

A. Caching Capacity Region

Let $G = (V_G, E_G)$ be an undirected capacitated graph, constructed as follows. G is a tree with leaf nodes $V(n) \subset V_G$. Leaf nodes in G share the same parent node in G if they fall within the same subsquare at level $L(n)$ in $A(n)$. Nodes at level ℓ in the tree G share the same parent node if all the leaf nodes that descend from it fall in the same subsquare at level $\ell - 1$ in $A(n)$. Note that through this construction, each set $V_{\ell,i}(n)$ for $\ell \in \{0, \ldots, L(n)\}$, $i \in 4^\ell$ is represented by exactly one internal node in G. This construction is illustrated in Figure 2. Assign to each edge $e \in E_G$ at level ℓ in G (i.e., between nodes at levels ℓ and $\ell - 1$) a capacity

$$c_e \triangleq \begin{cases} (4^{-\ell} n)^{2-\min\{3,\alpha\}/2} & \text{if } 1 \leq \ell \leq L(n), \\ 1 & \text{if } \ell = L(n) + 1. \end{cases}$$

With slight abuse of notation, we let for $(u,v) = e \in E_G$

$$c_{u,v} \triangleq c_e.$$

As we shall see in the following, the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ is closely related to the following quantity:

$$\widehat{\Lambda}^{\mathsf{CA}}(n) \triangleq \left\{ \lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n} : \sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}} \leq \sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v} \ \forall S \subset V_G \right\}.$$

The region $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is described by various subsets $S \subset V_G$. Each such subset can be understood as a *cut* in the graph G. For every cut $S \subset V_G$, the sum-rate

$$\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}}$$

between nodes in S and S^c (i.e., across the cut) is bounded by the sum-capacity

$$\sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v}$$

of edges between S and S^c. Note that we only count traffic $\lambda_{U,w}^{\mathsf{CA}}$ such that all caches U are contained in S.

The first result states that, for all $\alpha > 2$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is an approximate inner bound to the caching capacity region $\Lambda^{\mathsf{CA}}(n)$.

Theorem 1. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_1(n) \geq n^{-o(1)}$ such that*

$$b_1(n) \widehat{\Lambda}^{\mathsf{CA}}(n) \subset \Lambda^{\mathsf{CA}}(n)$$

with probability $1 - o(1)$ as $n \to \infty$.

We point out that Theorem 1 holds only with probability $1 - o(1)$ for different reasons in the fast and slow fading case. For fast fading, the theorem holds only for node placements that are "regular" enough. A random node placement satisfies these regularity conditions with high probability as $n \to \infty$. For slow fading, Theorem 1 holds under the same regularity conditions on the node placement, but moreover only holds for almost all realizations of the channel gains.

The next result states that, for all $\alpha > 6$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is also an approximate matching outer bound to $\Lambda^{\mathsf{CA}}(n)$.

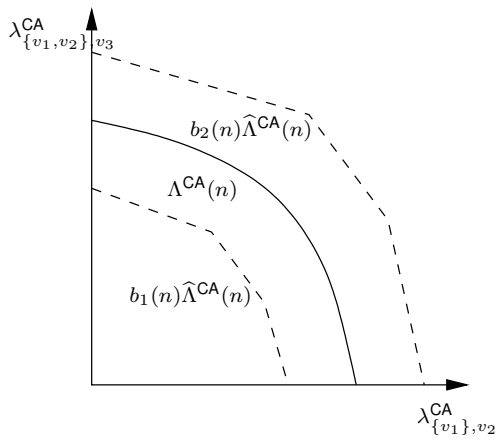

Fig. 3. For $\alpha > 6$, the set $\widehat{\Lambda}^{\mathsf{CA}}(n)$ approximates the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ of the wireless network in the sense that $b_1(n)\widehat{\Lambda}^{\mathsf{CA}}(n)$ (with $b_1(n) \geq n^{-o(1)}$) provides an inner bound to $\Lambda^{\mathsf{CA}}(n)$ and $b_2(n)\widehat{\Lambda}^{\mathsf{CA}}(n)$ (with $b_2(n) \leq n^{o(1)}$) provides an outer bound to $\widehat{\Lambda}^{\mathsf{CA}}(n)$. The figure shows two dimensions (namely $\lambda^{\mathsf{UC}}_{\{v_1\},v_2}$ and $\lambda^{\mathsf{UC}}_{\{v_1,v_2\},v_3}$) of the $2^n \times n$-dimensional sets $\Lambda^{\mathsf{CA}}(n)$ and $\widehat{\Lambda}^{\mathsf{CA}}(n)$.

Theorem 2. *Under either fast or slow fading, for any $\alpha > 6$, there exists $b_2(n) \leq n^{o(1)}$ such that*

$$\Lambda^{\mathsf{CA}}(n) \subset b_2(n)\widehat{\Lambda}^{\mathsf{CA}}(n)$$

with probability $1 - o(1)$ as $n \to \infty$.

As Theorem 1, Theorem 2 holds only with high probability due to regularity conditions on the node placement. However, unlike Theorem 1, Theorem 2 holds for all realizations of channel gains also for the slow fading case.

Comparing Theorems 1 and 2, we see that, for $\alpha > 6$, the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ is approximately equal to $\widehat{\Lambda}^{\mathsf{CA}}(n)$ in the sense that

$$n^{-o(1)}\widehat{\Lambda}^{\mathsf{CA}}(n) \subset \Lambda^{\mathsf{CA}}(n) \subset n^{o(1)}\widehat{\Lambda}^{\mathsf{CA}}(n).$$

In other words, for $\alpha > 6$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$ scales as the caching capacity region $\Lambda(n)$. This is illustrated in Figure 3.

B. Computational Aspects

Since we are interested in large networks, computational aspects are a concern. Note that the approximate caching capacity region $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is described in terms of essentially $\Theta(4^n)$ cuts

$S \subset V_G$. We show in Example 3 in Section III-D that a description with significantly fewer cuts is not possible. In other words, even an approximate description $\widehat{\Lambda}^{\mathsf{CA}}(n)$ of the caching capacity region $\Lambda^{\mathsf{CA}}(n)$ is computationally intractable for large values of n.

On the other hand, consider the simpler problem of testing membership of λ^{CA} in $\widehat{\Lambda}^{\mathsf{CA}}(n)$. We now argue that this problem can be approximately solved in an efficient manner. More precisely, we show that $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can be checked approximately in polynomial time in the description complexity of λ^{CA}. Combined with Theorems 1 and 2, this shows that, for $\alpha > 6$, approximate membership $\lambda^{\mathsf{CA}} \in \Lambda^{\mathsf{CA}}(n)$ can be checked efficiently as well.

Formally, define for any caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$

$$\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \triangleq \sup\{\rho \geq 0 : \rho\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)\}.$$

Membership $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can then be evaluated by checking if $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \leq 1$. Let $\phi_{\lambda^{\mathsf{CA}}}(n)$ to be the solution to the following linear program

$$\begin{array}{lrll} \max & \phi & & \\ \text{s.t.} & \sum\limits_{p \in P_{U,w}} f_{p,U,w} & \geq \phi\lambda^{\mathsf{CA}}_{U,w} & \forall\ U \subset V(n), w \in V(n), \\ \sum\limits_{p \in P: e \in p} \sum\limits_{U \subset V(n)} \sum\limits_{w \in V(n)} f_p & & \leq c_e & \forall\ e \in E_G, \\ & f_{p,U,w} & \geq 0 & \forall\ U \subset V(n), w \in V(n), p \in P_{U,w}, \end{array} \tag{1}$$

where $P_{u,w}$ is the path in G from node u to node w (since G is a tree, there is only one such path), and where

$$P_{U,w} \triangleq \bigcup_{u \in U} P_{u,w},$$

$$P \triangleq \bigcup_{U \subset V(n)} \bigcup_{w \in V(n)} P_{U,w}.$$

Note that the linear program (1), and hence also $\phi_{\lambda^{\mathsf{CA}}}(n)$, can be evaluated in polynomial time in the description length of λ^{CA} (i.e., in polynomial time in the length of the "input" of the linear program) by setting the flow variables $f_{p,U,w}$ to zero whenever $\lambda^{\mathsf{CA}}_{U,w} = 0$ and $p \in P_{U,w}$. Moreover, using a primal-dual algorithm, (1) can be solved efficiently in a distributed manner (see, for example, [22, Chapter 3.7]).

The following theorem shows that $\phi_{\lambda^{\mathsf{CA}}}(n)$ is a good approximation to $\rho_{\lambda^{\mathsf{CA}}}(n)$.

Theorem 3. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_3 \geq n^{-o(1)}$ such that for any n and caching traffic matrix $\lambda^{\mathsf{CA}} \mathbb{R}_+^{2^n \times n}$*

$$b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \leq \phi_{\lambda^{\mathsf{CA}}}(n) \leq \hat{\rho}_{\lambda^{\mathsf{CA}}}(n).$$

As argued above, $\phi_{\lambda^{\mathsf{CA}}}(n)$ can be computed in polynomial time in the description length of λ^{CA}. Hence Theorem 3 shows that testing membership $\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)$ can be done approximately in polynomial time in the description length of λ^{CA}. Combined with Theorems 1 and 2 this implies that, for $\alpha > 6$, approximate achievability of a traffic matrix λ^{CA} (i.e., testing membership $\lambda^{\mathsf{CA}} \in \Lambda^{\mathsf{CA}}(n)$) can be checked efficiently and in a distributed fashion.

C. An Efficient Content Delivery Protocol

Theorem 1 provides an inner bound to the caching capacity region of a wireless network. Here we describe the communication scheme achieving the inner bound. The matching outer bound shows that, for $\alpha > 6$, this scheme is optimal in the scaling sense.

Our proposed communication scheme consists of three layers, similar to a protocol stack. From high to low level of abstraction, these layers will be denoted by *routing layer*, *cooperation layer*, and *physical layer*.

From the view of the routing layer, the wireless network consists of the noiseless capacitated tree graph G defined in Section III-A (see Figure 2 there). To send a message available at the caches U to its destination w, the routing layer routes the message over G. The optimal requests of message parts from the caches in U (i.e., optimal cache selection) are found by solving the linear program (1). As pointed out in Section III-B, this optimal cache selection can be performed efficiently by a distributed algorithm.

The cooperation layer provides the tree abstraction G to the routing layer. Sending a message up or down an edge in the tree G in the routing layer corresponds in the cooperation layer to distributing or concentrating the same message in the wireless network. Recall that the leaf nodes of G are the nodes $V(n)$ of the wireless network and that each internal node of G represents some subsquare $V_{\ell,i}(n)$ of $V(n)$. To send a message from a child node to its parent in G (i.e., towards the root node of G), the message at the wireless nodes in $V(n)$ represented by the child node in G is distributed (over the wireless channel) evenly among all nodes in $V(n)$ represented by the parent node in G. This distribution is performed by splitting the message at each node

in $V(n)$ represented by the child note in G into equal size parts, and transmitting one part to each node in $V(n)$ represented by the parent node in G. To send a message from a parent node to a child node in G (i.e., away from the root node of G), the message at the wireless nodes in $V(n)$ represented by the parent node in G is concentrated on the wireless nodes in $V(n)$ represented by the child node in G. This concentration is performed be collecting at each node in $V(n)$ corresponding to the child node in G the message parts of the previously split up message located at the nodes in $V(n)$ corresponding to the parent node in G.

Finally, the physical layer performs this concentration or distribution of messages. Note that the kind of traffic resulting from the operation of the cooperation layer is regular in the sense that within each subsquare all nodes receive data at the same rate. Uniform traffic of this sort is well understood. Depending on the path-loss exponent α, we use either hierarchical cooperation [17], [18] (for $\alpha \in (2, 3]$) or multi-hop communication (for $\alpha > 3$). It is this operation of each edge in the physical layer that determines the edge capacity of the graph G as seen from the routing layer.

The next example illustrates the operation of this three-layer scheme. For more details on this architecture (in particular the cooperation and physical layers), we refer the reader to [19].

Example 2. Consider the three layers of the proposed communication architecture depicted in Figure 4. From top to bottom in the figure, these are the routing layer, the cooperation layer, and the physical layer. In this example, we consider a single (U, w) pair. The set of caches U consists of a single node $\{u\}$ in the wireless network shown at the bottom left, and its destination w is in the top right of the network. At the routing layer, the optimal choice of caches is in this case trivial (since there is just one cache u). The optimal route between u and w chosen at the routing layer is indicated in black dashed lines. Consider now the second edge along the path in G from u to w. The middle plane in the figure shows the induced behavior from using this edge in the cooperation layer. The bottom plane in the figure shows (part of) the corresponding actions induced in the physical layer. $\Diamond$

D. Example Scenarios

Here we provide three examples illustrating various aspects of the caching capacity region. Example 3 shows that the capacity region for caching is inherently more complicated than the

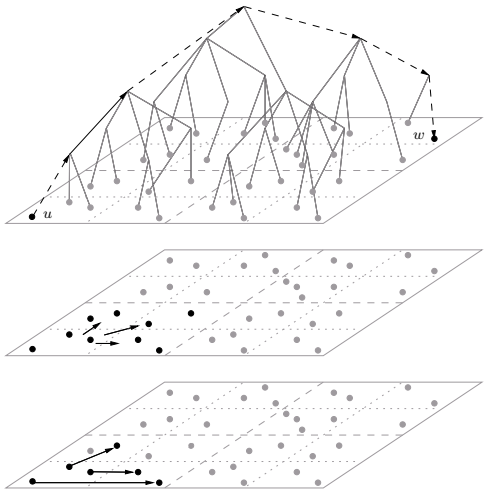

Fig. 4. Example operation of the three layer architecture.

ones resulting from unicast or multicast traffic. Example 4 shows that the strategy of always selecting the nearest cache can be arbitrarily bad. Example 5 analyzes the impact of complete caches on the performance of the wireless network.

Example 3. (*Insufficiency of edge cuts*)

For unicast traffic and multicast traffic, it is shown in [19] that it is sufficient to consider *edge cuts* in G, i.e, cuts that result if we remove a single edge from G. By construction, G has at most $2n$ edges, and hence there are at most $2n$ such edge cuts. This contrasts with the situation for caching traffic, for which Theorems 1 and 2 indicate that we have to consider general cuts, i.e., arbitrary subsets S of V_G. Indeed, the approximate capacity region $\widehat{\Lambda}^{\mathsf{CA}}(n)$ is expressed in terms of essentially $\Theta(4^n)$ cuts. Comparing these two results, one might suspect that a simpler characterization in terms of edge cuts can be found for the caching capacity region as well. This example shows that this is not possible. In other words, the caching capacity region is inherently more complicated than the unicast or multicast capacity regions of a wireless network.

Assume $V_{2,1}(n)$ and $V_{2,2}(n)$ are subsets of $V_{1,1}(n)$, and consider two nodes $u_1 \in V_{2,1}(n)$, $u_2 \in V_{2,2}(n)$. Construct

$$\lambda_{U,w}^{\mathsf{CA}} \triangleq \begin{cases} \rho(n) & \text{if } U = \{u_1, u_2\}, w \in V_{1,2}(n), \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. This is illustrated in Figure 5.

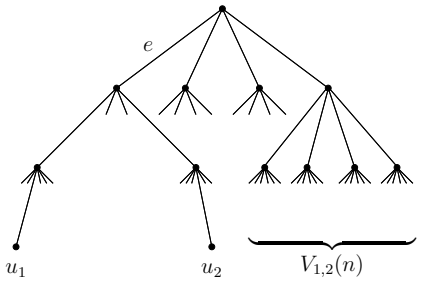

Fig. 5. Caching traffic pattern for Example 3.

The best edge cut results from removing edge e in Figure 5. The cut capacity is $c_e = \left(\frac{1}{4}n\right)^{2-\min\{3,\alpha\}/2}$ and the sum-rate across the cut is $|V_{1,2}(n)|\rho(n)$. By Theorem 2 and for $\alpha > 6$, this shows that $\rho^*(n)$, the largest achievable value of $\rho(n)$, is upper bounded as

$$\rho^*(n) \leq |V_{1,2}(n)|^{-1} n^{2-\min\{3,\alpha\}/2+o(1)} = n^{1-\min\{3,\alpha\}/2+o(1)}$$

with high probability.

On the other hand, consider the general node cut $S \triangleq \{u_1, u_2\} \subset V_G$. The cut capacity here is 2 and the sum-rate across the cut is again $|V_{1,2}(n)|\rho(n)$. Moreover, it is easily checked that S is the bottle neck cut in G. Thus, for $\alpha > 2$, Theorem 1 shows that $\rho^*(n)$ is lower bounded with high probability as

$$\rho^*(n) \geq |V_{1,2}(n)|^{-1} n^{-o(1)} = n^{-1-o(1)}, \tag{2}$$

and, for $\alpha > 6$, Theorem 2 shows that

$$\rho^*(n) \leq n^{-1+o(1)}.$$

In this example, it can be shown that the correct scaling of $\rho^*(n)$ is actually

$$\rho^*(n) = n^{-1\pm o(1)}$$

for all $\alpha > 2$ (not just $\alpha > 6$ as suggested by Theorem 2). Note that this differs substantially from the upper bound (2) obtained from the best edge cut. $\Diamond$

Example 4. (*Nearest-neighbor cache selection*)

A reasonable strategy of selecting caches is to request the entire message from the nearest available cache. In fact, this is the strategy implicitly assumed in most of the prior work considering caching in wireless networks cited in Section I-A. This example shows that this strategy can be arbitrarily bad.

Assume $V_{2,1}(n)$ and $V_{2,2}(n)$ are subsets of $V_{1,1}(n)$, and $V_{2,3}(n)$ is a subset of $V_{1,2}(n)$. Consider a node $u^* \in V_{2,2}(n)$, and label the nodes in $V_{2,1}(n) = \{w_1, w_2, \ldots\}$ and in $V_{2,3}(n) = \{u_1, u_2, \ldots\}$. Construct

$$\lambda_{U,w}^{\mathsf{CA}} \triangleq \begin{cases} \rho(n) & \text{if } U = \{u^*, u_i\}, w = w_i \text{ for some } i, \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. This is illustrated in Figure 6.

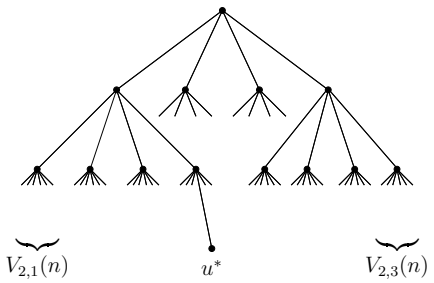

Fig. 6. Caching traffic pattern for Example 4.

For every w_i, the nearest cache is u^*. It can be shown that requesting the entire message from this nearest cache results in a per-node rate of at most

$$\rho(n) \leq n^{-1+o(1)}$$

for all $\alpha > 2$.

Assume now each w_i uses only the more distant cache u_i. This achieves a value of $\rho(n)$ of

$$\rho(n) \geq n^{1-\min\{3,\alpha\}/2-o(1)} \gg n^{-1+o(1)}.$$

Applying Theorem 1 yields the same $n^{1-\min\{3,\alpha\}/2-o(1)}$ value of $\rho(n)$, and Theorem 2 confirms that, for $\alpha > 6$, no scheme can achieve a better scaling. Hence

$$\rho^*(n) = n^{1-\min\{3,\alpha\}/2 \pm o(1)}$$

for $\alpha > 6$, and, as in the previous example, it can be shown that this is the correct scaling of $\rho^*(n)$ also for $\alpha \in (2, 6]$. This shows that the strategy of always selecting the nearest cache can result in a scaling exponent that is considerably worse than what is achievable with optimal cache selection. $\lozenge$

Example 5. (*Complete caches*)

Assume we randomly pick n^β caches for $\beta \in [0, 1)$, each holding a complete copy of all the messages. More precisely, letting $\widetilde{W} = \{w_i\}_{i=1}^{n^\beta}$ be the collection of caches, we consider a caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ of the form

$$\lambda_{W,v}^{\mathsf{CA}} = \begin{cases} \rho(n) & \text{if } W = \widetilde{W}, \\ 0 & \text{else,} \end{cases}$$

for some $\rho(n) \geq 0$. In this setup, choosing the nearest cache strategy (as discussed in Example 4) results in a per-node rate of

$$\rho(n) \geq n^{\beta - 1 - o(1)}$$

with probability $1 - o(1)$ as $n \to \infty$. The three-layer architecture proposed in Theorem 1 achieves the same rate, and Theorem 2 shows that, for $\alpha > 6$, for any communication scheme

$$\rho(n) \leq n^{\beta - 1 + o(1)}.$$

Hence, for $\alpha > 6$,

$$\rho^*(n) = n^{\beta - 1 \pm o(1)},$$

and it can be shown, as in the previous two examples, that this is the correct scaling of $\rho^*(n)$ also for $\alpha \in (2, 6]$.

This example illustrates that in situations in which the traffic demand and location of caches are regular enough, the strategy of selecting the nearest cache (as analyzed also in Example 4, and which is shown there to be arbitrarily bad in general) can actually be close to optimal. $\lozenge$

IV. Proofs

This section contains the proofs of Theorems 1, 2, and 3. We start in Section IV-A with some auxiliary results. Sections IV-B, IV-C, and IV-D contain the proofs of Theorems 3, 1, and 2, respectively.

A. Auxiliary Results

In this section, we define several quantities and recall some auxiliary results needed in several of the proofs.

We first introduce a "dual" description of the various regions. Recall that for any caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$

$$\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \triangleq \sup\left\{\rho \geq 0 : \rho\lambda^{\mathsf{CA}} \in \widehat{\Lambda}^{\mathsf{CA}}(n)\right\},$$

and define similarly

$$\rho_{\lambda^{\mathsf{CA}}}(n) \triangleq \sup\left\{\rho \geq 0 : \rho\lambda^{\mathsf{CA}} \in \Lambda^{\mathsf{CA}}(n)\right\}.$$

Consider a caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ for the wireless network and note that λ^{CA} can equivalently be treated as a traffic matrix between the leaf nodes of the graph G introduced in Section III. Let $\Lambda_G^{\mathsf{CA}}(n) \subset \mathbb{R}_+^{2^n \times n}$ be the collection of such caching traffic matrices $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ that can be routed over G. Note that $\phi_{\lambda^{\mathsf{CA}}}(n)$ as defined through the linear program (1) is equal to

$$\phi_{\lambda^{\mathsf{CA}}}(n) = \sup\left\{\phi \geq 0 : \phi\lambda^{\mathsf{CA}} \in \Lambda_G^{\mathsf{CA}}(n)\right\}.$$

It can be shown that the regions $\Lambda^{\mathsf{CA}}(n)$, $\widehat{\Lambda}^{\mathsf{CA}}(n)$, and $\Lambda_G^{\mathsf{CA}}(n)$ are convex, and hence knowledge of $\rho_{\lambda^{\mathsf{CA}}}(n)$, $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n)$, and $\phi_{\lambda^{\mathsf{CA}}}(n)$ for every $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$ is sufficient to completely describe them.

To analyze caching traffic in wireless networks, we shall make use of known results for unicast traffic in such networks. Formally, a *unicast traffic matrix* λ^{UC} for $V(n)$ is an element of $\mathbb{R}_+^{n \times n}$, associating with each pair $(u, w) \in V(n) \times V(n)$ the rate $\lambda_{u,w}^{\mathsf{UC}}$ at which node u wants to transmit a message to node w. We define the *unicast capacity region* $\Lambda^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n \times n}$ to be the collection all achievable unicast traffic matrices $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n}$. In analogy to the caching case, define $\Lambda_G^{\mathsf{UC}}(n) \subset \mathbb{R}_+^{n \times n}$ as the collection of unicast traffic matrices $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n}$ that can be routed over the tree graph G.

We now introduce some regularity conditions that are satisfied with high probability by a random node placement. Define $\mathcal{V}(n)$ to be the collection of all node placements $V(n)$ that satisfy the following conditions:

$$r_{u,v} > n^{-1} \qquad \text{for all } u, v \in V(n),$$

$$\big|V_{\ell,i}(n)\big| \leq \log(n) \qquad \text{for } \ell = \frac{1}{2}\log(n) \text{ and all } i \in \{1,\dots,4^\ell\},$$

$$\big|V_{\ell,i}(n)\big| \geq 1 \qquad \text{for } \ell = \frac{1}{2}\log\Big(\frac{n}{2\log(n)}\Big) \text{ and all } i \in \{1,\dots,4^\ell\},$$

$$\big|V_{\ell,i}(n)\big| \in [4^{-\ell-1}n, 4^{-\ell+1}n] \quad \text{for all } \ell \in \Big\{1,\dots,\frac{1}{2}\log(n)\big(1-\log^{-5/6}(n)\big)\Big\}, i \in \{1,\dots,4^\ell\}.$$

The first condition is that the minimum distance between node pairs is not too small. The second condition is that all squares of area 1 contain at most $\log(n)$ nodes. The third condition is that all squares of area $2\log(n)$ contain at least one node. The fourth condition is that all squares up to level $\frac{1}{2}\log(n)\big(1-\log^{-5/6}(n)\big)$ contain a number of nodes proportional to their area.

The next lemma states that a random node placement satisfies these conditions with high probability.

Lemma 4.

$$\mathbb{P}\big(V(n) \in \mathcal{V}(n)\big) \geq 1 - o(1)$$

as $n \to \infty$.

Proof. See [19, Lemma 5]. □

B. Proof of Theorem 3

We first prove the upper bound, i.e.,

$$\phi_{\lambda^{\mathsf{CA}}} \leq \hat{\rho}_{\lambda^{\mathsf{CA}}}. \tag{3}$$

Note that if $\lambda^{\mathsf{CA}} \in \Lambda_G^{\mathsf{CA}}$ then there exists a strategy to route traffic at rates λ^{CA} over G. This implies that the flow across each cut $S \subset V_G$ must be less than the capacity of that cut. The flow across such a cut S contains at least all those requested messages that only contain caches in S, i.e.,

$$\sum_{U \subset S \cap V(n)} \sum_{w \in V(n) \setminus S} \lambda_{U,w}^{\mathsf{CA}}.$$

On the other hand, the capacity of the cut S is equal to

$$\sum_{\substack{(u,v) \in E_G: \\ u \in S, v \notin S}} c_{u,v}.$$

Therefore

$$\sum_{U\subset S\cap V(n)}\sum_{w\in V(n)\setminus S}\lambda^{\mathsf{CA}}_{U,w}\le\sum_{\substack{(u,v)\in E_G:\\ u\in S,v\notin S}}c_{u,v}$$

for all $S\subset V$, and hence $\lambda^{\mathsf{CA}}\in\Lambda^{\mathsf{CA}}_G$ implies $\lambda^{\mathsf{CA}}\in\widehat{\Lambda}^{\mathsf{CA}}$. Thus, $\Lambda^{\mathsf{CA}}_G\subset\widehat{\Lambda}^{\mathsf{CA}}$, from which (3) follows.

We now prove the lower bound, i.e., we show that there exists $b_3(n)\ge n^{-o(1)}$ such that for any λ^{CA}

$$\phi_{\lambda^{\mathsf{CA}}}\ge b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}. \tag{4}$$

Pick any λ^{CA}. Since for any $b>0$,

$$\phi_{b\lambda^{\mathsf{CA}}}=\frac{1}{b}\phi_{\lambda^{\mathsf{CA}}},$$
$$\hat{\rho}_{b\lambda^{\mathsf{CA}}}=\frac{1}{b}\hat{\rho}_{\lambda^{\mathsf{CA}}},$$

we may assume without loss of generality that

$$\sum_{(U,w)}\lambda^{\mathsf{CA}}_{U,w}=1. \tag{5}$$

Recall that G is an *undirected* capacitated graph. Construct a *directed* capacitated graph $\widetilde{G}=(V_{\widetilde{G}},E_{\widetilde{G}})$ as follows. Take the undirected graph G and turn it into a directed graph by splitting each edge $e\in E_G$ into two directed edges each with the same capacity as e. Add 2^n additional nodes to V_G, one for each subset $U\subset V$. Connect the new node $\tilde{u}$ corresponding to $U\subset V$ to each node $u\in U$ by a (directed) edge $(\tilde{u},u)$ with $c_{\tilde{u},u}=\infty$. This procedure is illustrated in Figure 7. We call the directed version of G that is contained in $\widetilde{G}$ as a subgraph its *core*. Note that if some flows can be routed through G then the same flows can be routed through the core of $\widetilde{G}$, and if some flows can routed through the core of $\widetilde{G}$ then at least half of each flow can be routed through G. Hence, for scaling purposes, the two are equivalent.

Now, assume we are given a caching traffic matrix λ^{CA} for G. Construct a *unicast* traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ for $\widetilde{G}$ by making for each (U,w) pair in G (i.e., $U\subset V$, $w\in V$) the node $\tilde{u}$ in $\widetilde{G}$ corresponding to U a source for w with rate

$$\tilde{\lambda}^{\mathsf{UC}}_{\tilde{u},w}\triangleq\lambda^{\mathsf{CA}}_{U,w}.$$

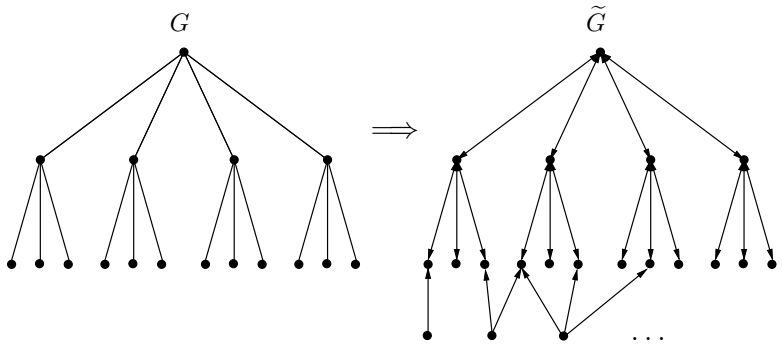

Fig. 7. Construction of the directed graph $\widetilde{G}$ from the undirected graph G.

Denote by $\Lambda_{\widetilde{G}}^{\mathsf{UC}}$ the set of feasible unicast traffic matrices for $\widetilde{G}$, and set

$$\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} \triangleq \sup\left\{\phi \geq 0 : \phi\tilde{\lambda}^{\mathsf{UC}} \in \Lambda_{\widetilde{G}}^{\mathsf{UC}}\right\}.$$

By construction of $\widetilde{G}$ from G, and by the above argument relating G to the core of $\widetilde{G}$, we have

$$\phi_{\lambda^{\mathsf{CA}}} \geq \frac{1}{2}\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}}. \tag{6}$$

We are thus left with the problem of analyzing unicast traffic over $\widetilde{G}$. Two difficulties arise. First, $\widetilde{G}$ is a directed graph. While unicast traffic over undirected graphs with m nodes are well understood and $O(\log(m))$ approximation results for the capacity region of such graphs in terms of cut-set bounds are known [23], the best known approximation result for general directed graphs is (up to polylog factors) $O(m^{11/23})$ [24]. Second, the graph $\widetilde{G}$ is exponentially big in n. More precisely, $|V_{\widetilde{G}}| \geq 2^n$. Hence even a logarithmic (in the size m of the graph) approximation result will only yield a polynomial approximation in n. Nonetheless, as we shall see, the special structure of $\widetilde{G}$ can be used to obtain $\log(n)$ approximation results of $\Lambda_{\widetilde{G}}^{\mathsf{UC}}$.

We use an idea from [25], namely that the unicast traffic problem can be reduced to a maximum sum-rate problem. More precisely, for a subset $\widetilde{F} \subset V_{\widetilde{G}} \times V_{\widetilde{G}}$ of (u, w) pairs in $\widetilde{G}$, define the *maximum sum-rate* as

$$\tilde{\sigma}_{\widetilde{F}} \triangleq \sup\left\{\textstyle\sum_{(u,w)\in\widetilde{F}} \tilde{\lambda}_{u,w}^{\mathsf{UC}} : \tilde{\lambda}^{\mathsf{UC}} \in \Lambda_{\widetilde{G}}^{\mathsf{UC}}\right\}.$$

We now argue that for every unicast traffic matrix $\tilde{\lambda}^{\mathsf{UC}}$ there exists $\widetilde{F}$ such that $\tilde{\sigma}_{\widetilde{F}}$ is not too much bigger than $\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}}$.

First, note that $\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}}$ is the solution to the following linear program

$$\begin{array}{lrll} \text{maximize} & \phi & & \\ \text{subject to} & \sum_{p\in\widetilde{P}_{u,w}} f_p & \geq \phi\tilde{\lambda}^{\mathsf{UC}}_{u,w} & \forall\ u,w\in V_{\widetilde{G}}, \\ & \sum_{p\in\widetilde{P}:e\in p} f_p & \leq c_e & \forall e\in E_{\widetilde{G}}, \\ & f_p & \geq 0 & \forall\ p\in\widetilde{P}, \end{array} \tag{7}$$

where $\widetilde{P}_{u,w}$ is the collection of all paths in $\widetilde{G}$ from node u to node w, and

$$\tilde{P} \triangleq \bigcup_{(u,w)\in V_{\widetilde{G}}\times V_{\widetilde{G}}} \widetilde{P}_{u,w}.$$

The corresponding dual linear program is

$$\begin{array}{lrll} \text{minimize} & \sum_{e\in E_{\widetilde{G}}} c_e m_e & & \\ \text{subject to} & \sum_{e\in p} m_e & \geq d_{u,w} & \forall\ u,w\in V_{\widetilde{G}}, p\in\widetilde{P}_{u,w}, \\ & \sum_{u,w\in V_{\widetilde{G}}} d_{u,w}\tilde{\lambda}^{\mathsf{UC}}_{u,w} & \geq 1 & \\ & m_e & \geq 0 & \forall\ e\in E_{\widetilde{G}}, \\ & d_{u,w} & \geq 0 & \forall\ u,w\in V_{\widetilde{G}}. \end{array} \tag{8}$$

Since the all-zero solution is feasible for the primal program (7), strong duality holds.

Second, $\tilde{\sigma}_{\widetilde{F}}$ is the solution to the linear program

$$\begin{array}{lrll} \text{maximize} & \sum_{(u,w)\in\widetilde{F}}\sum_{p\in\widetilde{P}_{u,w}} f_p & & \\ \text{subject to} & \sum_{p\in\widetilde{P}:e\in p} f_p & \leq c_e & \forall e\in E_{\widetilde{G}}, \\ & f_p & \geq 0 & \forall\ p\in\widetilde{P}, \end{array}$$

and its dual is

$$\begin{array}{lrll} \text{minimize} & \sum_{e\in E_{\widetilde{G}}} c_e m_e & & \\ \text{subject to} & \sum_{e\in p} m_e & \geq d_{u,w} & \forall\ u,w\in V_{\widetilde{G}}, p\in\widetilde{P}_{u,w}, \\ & d_{u,w} & \geq 1 & \forall\ (u,w)\in\widetilde{F}, \\ & m_e & \geq 0 & \forall\ e\in E_{\widetilde{G}}, \\ & d_{u,w} & \geq 0 & \forall\ u,w\in V_{\widetilde{G}}. \end{array} \tag{9}$$

Again strong duality holds.

Let $\{m^*_e\}_{e\in E_{\widetilde{G}}}$, $\{d^*_{u,w}\}_{u,w\in V_{\widetilde{G}}}$ be a minimizer for the dual (8) of the unicast traffic problem. We now show how $\{m^*_e\}$, $\{d^*_{u,w}\}$ can be used to construct a solution to the dual (9) of the maximum

sum-rate problem. Note that we can assume without loss of optimality that

$$d^*_{u,w} = \begin{cases} 0 & \text{if } \tilde{\lambda}^{\mathsf{UC}}_{u,w} = 0, \\ \min_{p \in \widetilde{P}_{u,w}} \sum_{e \in p} m^*_e & \text{else.} \end{cases} \tag{10}$$

Now, since $c_e = \infty$ whenever $e \in E_{\widetilde{G}} \setminus E_G$, we have $m^*_e = 0$ for those edges. Since, in addition, $\tilde{\lambda}^{\mathsf{UC}}_{u,w} > 0$ only if $u \in V_{\widetilde{G}} \setminus V_G$ and if w is a leaf node of G, this implies that $\{d^*_{u,w}\}_{u,w \in V_{\widetilde{G}}}$ can take at most n^2 different nonzero values. Order these values in decreasing order

$$d^*_1 > d^*_2 > \ldots > d^*_K > d^*_{K+1} = 0$$

with $K \leq n^2$, and define

$$\tilde{\lambda}^{\mathsf{UC}}_k \triangleq \sum_{u,w \in V_{\widetilde{G}} : d^*_{u,w} = d^*_k} \tilde{\lambda}^{\mathsf{UC}}_{u,w}.$$

We now argue that $d^*_k \leq n^2$ for all $k \in \{1, \ldots, K\}$. In fact, assume $d^*_1 > n^2$, then by (10) there exists at least one edge $\tilde{e}$ such that $m^*_{\tilde{e}} > n$. Hence

$$\sum_{e \in E_{\widetilde{G}}} c_e m^*_e \geq c_{\tilde{e}} m^*_{\tilde{e}} > n$$

since $c_e \geq 1$ for all $e \in E_{\widetilde{G}}$. On the other hand, let $m_e = 1$ for all edges between the leave nodes and parent nodes in the core of $\widetilde{G}$, and let $m_e = 0$ for all other edges. Set $d_{u,w}$ as in (10) but with respect to this choice of $\{m_e\}$. Since all paths between node pairs (u, w) such that $\tilde{\lambda}^{\mathsf{UC}}_{u,w} > 0$ include at least one edge between the aforementioned leave and parent nodes, we have $d_{u,w} \geq 1$ whenever $\tilde{\lambda}^{\mathsf{UC}}_{u,w} > 0$, and therefore

$$\sum_{u,w \in V_{\widetilde{G}}} d_{u,w} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \geq \sum_{u,w \in V_{\widetilde{G}}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} = 1,$$

by the normalization assumption (5). Thus $\{m_e\}$, $\{d_{u,v}\}$ is feasible for the dual (8), and has value

$$\sum_{e \in E_{\widetilde{G}}} c_e m_e = n < \sum_{e \in E_{\widetilde{G}}} c_e m^*_e,$$

contradicting the optimality of $\{m^*_e\}$, $\{d^*_{u,v}\}$. Hence $d^*_k \leq d^*_1 \leq n^2$ for all k.

We now argue that at least one d^*_k is not too small. Let $k_1 < k_2 < \ldots < k_I$ be such that

$$\{k_i\}_{i=1}^I = \Big\{ k : \tilde{\lambda}^{\mathsf{UC}}_k \geq \frac{1}{2n^4} \Big\}. \tag{11}$$

Note that $I \geq 1$ since otherwise

$$\begin{aligned}\sum_{u,w\in V_{\widetilde{G}}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} &= \sum_{k=1}^{K+1} \tilde{\lambda}^{\mathsf{UC}}_{k} \\ &< (K+1)\frac{1}{2n^4} \\ &\leq \frac{n^2+1}{2n^4} \\ &\leq 1,\end{aligned}$$

contradicting the normalization assumption (5). Finally, define

$$s_i \triangleq \sum_{j=1}^{i} \tilde{\lambda}^{\mathsf{UC}}_{k_j}.$$

Using that $\{d^*_k\}$ is feasible for the dual (8), that $d^*_k \leq n^2$, and that $K \leq n^2$, we have

$$\begin{aligned}\sum_{i=1}^{I} d^*_{k_i}\tilde{\lambda}^{\mathsf{UC}}_{k_i} &\geq 1 - \sum_{k:\tilde{\lambda}^{\mathsf{UC}}_k < 1/2n^4} d^*_k\tilde{\lambda}^{\mathsf{UC}}_{k} \\ &\geq 1 - \frac{1}{2n^4}Kn^2 \\ &\geq \frac{1}{2}. \end{aligned} \tag{12}$$

We argue that this implies existence of i such that

$$d^*_{k_i} \geq \frac{1}{2s_i(1+\ln(2n^4))}. \tag{13}$$

Indeed, assume (13) is false for all i. Then

$$\begin{aligned}\sum_{i=1}^{I} d^*_{k_i}\tilde{\lambda}^{\mathsf{UC}}_{k_i} &< \frac{1}{2(1+\ln(2n^4))}\sum_{i=1}^{I}\frac{\tilde{\lambda}^{\mathsf{UC}}_{k_i}}{s_i} \\ &= \frac{1}{2(1+\ln(2n^4))}\Big(1+\sum_{i=2}^{I}\frac{s_i - s_{i-1}}{s_i}\Big) &\text{(14a)} \\ &\leq \frac{1}{2(1+\ln(2n^4))}\Big(1+\sum_{i=2}^{I}\big(\ln(s_i)-\ln(s_{i-1})\big)\Big) &\text{(14b)} \\ &= \frac{1}{2(1+\ln(2n^4))}\big(1+\ln(s_I/\tilde{\lambda}^{\mathsf{UC}}_{k_1})\big) \\ &\leq \frac{1}{2(1+\ln(2n^4))}\big(1+\ln(2n^4)\big) &\text{(14c)} \\ &= \frac{1}{2},\end{aligned}$$

where we have used that $I \geq 1$ in (14a), that $1 - x \leq -\ln(x)$ for every $x \geq 0$ in (14b), and that $s_I \leq 1$ by (5) and $\tilde{\lambda}_{k_1}^{\mathsf{UC}} \geq \frac{1}{2n^4}$ in (14c). This contradicts (12), showing that (13) must hold for some i. Consider this value of i in the following.

Now, consider the following set $\widetilde{F}$ of (u, w) pairs:

$$\widetilde{F} \triangleq \left\{(u, w) : d^*_{u,w} \geq d^*_{k_i}\right\}.$$

Note that, by (10), $\widetilde{F}$ contains only pairs (u, w) such that $u \in V_{\widetilde{G}} \setminus V_G$ and $w \in V \subset V_{\widetilde{G}}$ (i.e., nodes in $\widetilde{G}$ corresponding to leaf nodes in G). Set

$$d_{u,w} \triangleq \frac{d^*_{u,w}}{d^*_{k_i}},$$
$$m_e \triangleq \frac{m^*_e}{d^*_{k_i}}.$$

Note that for $(u, w) \in \widetilde{F}$,

$$d_{u,w} = \frac{d^*_{u,w}}{d^*_{k_i}} \geq 1,$$

and that for all $u, w \in V_{\widetilde{G}}$, $p \in \widetilde{P}_{u,w}$

$$\begin{aligned}\sum_{e \in p} m_e &= \frac{1}{d^*_{k_i}} \sum_{e \in p} m^*_e \\ &\geq \frac{1}{d^*_{k_i}} d^*_{u,w} \\ &= d_{u,w},\end{aligned}$$

by feasibility of $\{d^*_{u,w}\}$ and $\{m^*_e\}$ for the dual (8). Hence, for this $\widetilde{F}$, the choice of $\{m_e\}$ and $\{d_{u,w}\}$ is feasible for the dual (9). By weak duality

$$\begin{aligned}\tilde{\sigma}_{\widetilde{F}} &\leq \sum_{e \in E_{\widetilde{G}}} c_e m_e \\ &= \frac{1}{d^*_{k_i}} \sum_{e \in E_{\widetilde{G}}} c_e m^*_e.\end{aligned}$$

By (13),

$$d^*_{k_i} \geq \frac{1}{2 s_i (1 + \ln(2n^4))},$$

and, since $d^*_{k_j} \geq d^*_{k_i}$ for all $j \leq i$,

$$\begin{aligned} s_i &= \sum_{j=1}^{i} \tilde{\lambda}^{\mathsf{UC}}_{k_j} \\ &= \sum_{j=1}^{i} \sum_{(u,w):d^*_{u,w}=d_{k^*_j}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\ &\leq \sum_{(u,w):d^*_{u,w}\geq d^*_{k_i}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\ &= \sum_{(u,w)\in\widetilde{F}} \tilde{\lambda}^{\mathsf{UC}}_{u,w} \\ &\triangleq \tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}. \end{aligned}$$

Therefore

$$\tilde{\sigma}_{\widetilde{F}} \leq 2\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}(1+\ln(2n^4)) \sum_{e\in E_{\widetilde{G}}} c_e m^*_e.$$

Since $\{m^*_e\}$ is optimal for the dual (8), and by strong duality, we also have

$$\sum_{e\in E_{\widetilde{G}}} c_e m^*_e = \tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}},$$

and hence

$$\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} \geq \frac{1}{2(1+\ln(2n^4))} \frac{\tilde{\sigma}_{\widetilde{F}}}{\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}}. \tag{15}$$

We are thus left with analyzing maximum sum-rates $\tilde{\sigma}_{\widetilde{F}}$ in $\widetilde{G}$. Now notice that, since the edges in $E_{\widetilde{G}} \setminus E_G$ have infinite capacity, and since for $(u,w) \in \widetilde{F}$ we have $u \in V_{\widetilde{G}} \setminus V_G$ and $w \in V \subset V_G \subset V_{\widetilde{G}}$, this analysis can be done by considering only the core of $\widetilde{G}$. More precisely, for a collection of node pairs $\widetilde{F}$ in $\widetilde{G}$ as above, we construct a collection of node pairs F in G as follows. For each $(\tilde{u},w) \in \widetilde{F}$ with $\tilde{u}$ connected by $\widetilde{G}$ with nodes $U \subset V_G \subset V_{\widetilde{G}}$, add (u,w) to F for each $u \in U$. Denote by σ_F the maximum sum-rate for F in G. Since G is the undirected version of the core of $\widetilde{G}$, we have

$$\tilde{\sigma}_{\widetilde{F}} \geq \sigma_F. \tag{16}$$

For a collection of node pairs F in G, we call a set of edges M a *multicut* for F if in the graph $(V_G, E_G \setminus M)$ each pair in F is disconnected. For a subset $M \subset E_G$, define

$$c_M \triangleq \sum_{e\in M} c_e.$$

It is shown in [26, Theorem 8] that if G is an undirected tree, then for every $F \in V_G \times V_G$ there exists a multicut M for F such that

$$\sigma_F \geq \frac{1}{2} c_M. \tag{17}$$

Combining (15), (16), and (17), we obtain that for every $\tilde{\lambda}^{\mathsf{UC}}$ there exists a collection of node pairs $\widetilde{F}$ in $\widetilde{G}$, and a multicut M for the corresponding F in G such that

$$\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} \geq \frac{1}{4(1+\ln(2n^4))} \frac{c_M}{\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}}}. \tag{18}$$

We now show how the edge cut $M \subset E_G$ can be transformed into a node cut $S \subset V_G$. Denote by $\{S_i\}$ the connected components of $(V_G, E_G \setminus M)$. We can assume without loss of generality that

$$M = \bigcup_i (S_i \times S_i^c) \cap E_G,$$

since otherwise we can remove the additional edges from M to create a smaller multicut for F. We therefore have

$$c_M = \frac{1}{2} \sum_i c_{(S_i^c \times S_i) \cap E_G}, \tag{19}$$

since every edge in M appears exactly twice in the sum on the right-hand side. Define for $S \subset V_G$

$$\lambda^{\mathsf{CA}}_{S,S^c} \triangleq \sum_{U \subset S \cap V} \sum_{w \in V \setminus S} \lambda^{\mathsf{CA}}_{U,w}.$$

M is a multicut for F induced by $\widetilde{F}$, and hence for every $(u,w) \in \widetilde{F}$ and the corresponding pair (U,w), M separates w from all the nodes in U. Therefore, for each such (U,w) pair, there exists a S_i such that $w \in S_i$, $U \subset S_i^c$. This shows that

$$\tilde{\lambda}^{\mathsf{UC}}_{\widetilde{F}} \leq \sum_i \lambda^{\mathsf{CA}}_{S_i^c, S_i}. \tag{20}$$

Equations (18), (19), and (20) imply that there exists j such that

$$\begin{aligned}
\tilde{\phi}_{\tilde{\lambda}^{\mathsf{UC}}} &\geq \frac{1}{8(1+\ln(2n^4))} \frac{\sum_i c_{(S_i^c \times S_i) \cap E_G}}{\sum_i \lambda^{\mathsf{CA}}_{S_i^c, S_i}} \\
&\geq \frac{1}{8(1+\ln(2n^4))} \frac{c_{(S_j^c \times S_j) \cap E_G}}{\lambda^{\mathsf{CA}}_{S_j^c, S_j}} \\
&\geq \frac{1}{8(1+\ln(2n^4))} \min_{S \subset V_G} \frac{c_{(S \times S^c) \cap E_G}}{\lambda^{\mathsf{CA}}_{S,S^c}} \\
&= \frac{1}{8(1+\ln(2n^4))} \hat{\rho}_{\lambda^{\mathsf{CA}}}.
\end{aligned} \tag{21}$$

Combined with (6), this shows that for

$$b_3(n) \triangleq \frac{1}{16(1+\ln(2n^4))} \geq n^{-o(1)}$$

we have

$$\phi_{\lambda^{\mathsf{CA}}} \geq b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}},$$

proving the lower bound in Theorem 3.

C. Proof of Theorem 1

In this Section, we provide the proof of Theorem 1. Instead of proving the theorem directly, it will be convenient to work with the dual descriptions $\rho_{\lambda^{\mathsf{CA}}}(n)$ and $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n)$ of $\Lambda^{\mathsf{CA}}(n)$ and $\widehat{\Lambda}^{\mathsf{CA}}(n)$ introduced in Section IV-A. The next theorem is the dual version of Theorem 1.

Theorem 5. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_1 = n^{-o(1)}$ such that with probability $1 - o(1)$ as $n \to \infty$ for any n and caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$*

$$b_1(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}(n) \leq \rho_{\lambda^{\mathsf{CA}}}(n).$$

Proof. The same arguments as in [19, Theorem 1] show that there exists $b(n) \geq n^{-o(1)}$ such that if a caching traffic matrix λ^{CA} can be routed over G, then $b(n)\lambda^{\mathsf{CA}}$ can be communicated reliably over the wireless network. Formally, if $V \in \mathcal{V}$ then under fast fading

$$b(n)\phi_{\lambda^{\mathsf{CA}}} \leq \rho_{\lambda^{\mathsf{CA}}}, \tag{22}$$

and the same results holds for slow fading for a collection of channel gains $\mathcal{H}$ (not dependent on λ^{CA}) with

$$\mathbb{P}(\{h_{u,v}\}_{u,v\in V} \in \mathcal{H}) \geq 1 - o(1)$$

as $n \to \infty$.

Combining (22), with Theorem 3 and Lemma 4, we obtain that with probability

$$\mathbb{P}(\{h_{u,v}\} \in \mathcal{H}, V \in \mathcal{V}) \geq 1 - o(1)$$

as $n \to \infty$, we have for any caching traffic matrix λ^{CA}

$$\begin{aligned}\rho_{\lambda^{\mathsf{CA}}} &\geq b(n)\phi_{\lambda^{\mathsf{CA}}} \\ &\geq b(n)b_3(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}.\end{aligned}$$

Setting

$$b_1(n) \triangleq b(n)b_3(n),$$

and recalling that $b_3(n) \geq n^{-o(1)}$ and $b(n) \geq n^{-o(1)}$ both uniformly in λ^{CA}, concludes the proof of Theorem 5. □

D. Proof of Theorem 2

In this Section, we prove Theorem 2. As before, it will be convenient to work with the dual description $\rho_{\lambda^{\mathsf{CA}}}(n)$ and $\hat{\rho}_{\lambda^{\mathsf{CA}}}(n)$ of $\Lambda^{\mathsf{CA}}(n)$ and $\widehat{\Lambda}^{\mathsf{CA}}(n)$ as introduced in Section IV-A. The next theorem is the dual version of Theorem 2

Theorem 6. *Under either fast or slow fading, for any $\alpha > 2$, there exists $b_2 \leq n^{o(1)}$ such that with probability $1 - o(1)$ as $n \to \infty$ for any n and caching traffic matrix $\lambda^{\mathsf{CA}} \in \mathbb{R}_+^{2^n \times n}$*

$$\rho_{\lambda^{\mathsf{CA}}}(n) \leq b_2(n)\hat{\rho}_{\lambda^{\mathsf{CA}}}(n).$$

We start with some auxiliary lemmas. For a subsets $S_1, S_2 \subset V(n)$, denote by $C(S_1, S_2)$ the MIMO capacity between the nodes in S_1 and S_2. Denote by S_2^k the nodes in S_2 that are at distance between k and $k+1$ from S_1, i.e.,

$$S_2^k \triangleq \{v \in S_2 : \min_{u \in S_1} r_{u,v} \in [k, k+1)\}.$$

Lemma 7. *Under either fast or slow fading, for every $\alpha > 6$, there exists a constant K_1 such that for all $V(n) \in \mathcal{V}(n)$ and all $S \subset V(n)$*

$$C(S, S^c) \leq K_1 \log^4(n) \sum_{k=0}^{\log(n)} |S_2^k|.$$

Proof. Set $S_1 \triangleq S$ and $S_2 \triangleq S^c$, and note that

$$S_2 = \bigcup_{k=0}^{\infty} S_2^k,$$

Let

$$\mathbf{H}_{S_1,S_2} \triangleq [h_{u,v}]_{u \in S_1, v \in S_2}$$

be the matrix of channel gains between the nodes in S_1 and S_2. Under fast fading

$$C(S_1, S_2) \triangleq \max_{\substack{\mathbf{Q}(\mathbf{H}) \geq 0: \\ \mathbb{E}(q_{u,u}) \leq P \ \forall u \in S_1}} \mathbb{E}\bigg(\log\det\left(\mathbf{I} + \mathbf{H}_{S_1,S_2}^{\dagger}\mathbf{Q}(\mathbf{H})\mathbf{H}_{S_1,S_2}\right)\bigg),$$

and under slow fading

$$C(S_1, S_2) \triangleq \max_{\substack{\mathbf{Q}\geq 0: \\ q_{u,u}\leq P \ \forall u\in S_1}} \log\det\left(\mathbf{I} + \mathbf{H}^*_{S_1,S_2}\mathbf{Q}\mathbf{H}_{S_1,S_2}\right).$$

Applying the generalized Hadamard inequality, we obtain that under either fast or slow fading

$$C(S_1, S_2) \leq C(S_1, \cup_{k=0}^{\log(n)} S_2^k) + C(S_1, \cup_{k>\log(n)} S_2^k). \tag{23}$$

Now, for the first term in (23), using Hadamard's inequality once more, yields

$$\begin{aligned} C(S_1, \cup_{k=0}^{\log(n)} S_2^k) &\leq \sum_{k=0}^{\log(n)} \sum_{v\in S_2^k} C(S_1, \{v\}) \\ &\leq \sum_{k=0}^{\log(n)} \sum_{v\in S_2^k} C(\{v\}^c, \{v\}). \end{aligned}$$

By Lemma 7 in [19],

$$C(\{v\}^c, \{v\}) \leq K\log(n)$$

for some constant K, and thus

$$C(S_1, \cup_{k=0}^{\log(n)} S_2^k) \leq K\log(n) \sum_{k=0}^{\log(n)} |S_2^k|. \tag{24}$$

For the second term in (23), we have the following upper bound from (slightly adapting) Theorem 2.1 in [11]:

$$C(S_1, \cup_{k>\log(n)} S_2^k) \leq \sum_{k>\log(n)} \sum_{v\in S_2^k} \Big(\sum_{u\in S_1} r_{u,v}^{-\alpha/2}\Big)^2.$$

For $v \in S_2^k$, the (open) disk of radius k around v does not contain any node in S_1 (by definition of S_2^k). Moreover, since $V \in \mathcal{V}$, there are at most $\log(n)$ nodes inside every subsquare of A of sidelength one. Thus

$$\begin{aligned} \sum_{u\in S_1} r_{u,v}^{-\alpha/2} &\leq \log(n) \sum_{\tilde{k}=k}^{\infty} 8\pi(\tilde{k}+2)\tilde{k}^{-\alpha/2} \\ &\leq \widetilde{K}\log(n)k^{2-\alpha/2}, \end{aligned}$$

for some constant $\widetilde{K}$ independent of S_1 and k. Therefore,

$$C(S_1, \cup_{k>\log(n)} S_2^k) \leq \sum_{k>\log(n)} |S_2^k| \widetilde{K}^2 \log^2(n) k^{4-\alpha}. \tag{25}$$

Consider now some $v \in S_2^k$ with $k > \log(n)$, and let u^* be the closest point in S_1 to v. Since $v \in S_2^k$, we must have

$$r_{u^*,v} \in [k, k+1).$$

Consider the (open) disk of radius $r_{u^*,v}$ around v and the disk of radius $\log(n)$ around u^*. Since u^* is the closest node to v in S_1, all nodes in the disk around v are in S_2. Moreover, the intersection of the two disks has an area of at least $\frac{\pi}{4}\log^2(n)$. Since $V \in \mathcal{V}$, this implies that this intersection must contain at least one point, say $\tilde{v}$, and by construction

$$\tilde{v} \in \bigcup_{\tilde{k}=0}^{\log(n)} S_2^{\tilde{k}}.$$

This shows that for every node v in S_2^k there exists a node $\tilde{v}$ in $\cup_{\tilde{k}=0}^{\log(n)} S_2^{\tilde{k}}$ such that

$$r_{v,\tilde{v}} \in [k - \log(n), k+1).$$

Now, since $V \in \mathcal{V}$, for every node $\tilde{v}$, there are at most

$$2\pi(k+1)(\log(n)+5)\log(n) \leq K' k \log^2(n)$$

nodes at distance $[k - \log(n), k+1)$. Hence the number of nodes in S_2^k is at most

$$|S_2^k| \leq K' k \log^2(n) \sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}|. \tag{26}$$

Combining (26) with (25) yields

$$\begin{aligned} C(S_1, \cup_{k>\log(n)} S_2^k) &\leq \sum_{k>\log(n)} |S_2^k| \widetilde{K}^2 \log^2(n) k^{4-\alpha} \\ &\leq K' \widetilde{K}^2 \log^4(n) \Big(\sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}| \Big) \sum_{k>\log(n)} k^{5-\alpha} \\ &= K'' \log^4(n) \sum_{\tilde{k}=0}^{\log(n)} |S_2^{\tilde{k}}|, \end{aligned} \tag{27}$$

for some constant K'', and where we have used that $\alpha > 6$. Finally, plugging (24) and (27) into (23) shows that

$$C(S_1, S_2) \leq (K + K'') \log^4(n) \sum_{k=0}^{\log(n)} |S_2^k|,$$

which proves the lemma with

$$K_1 \triangleq K + K''. \qquad \square$$

The next lemma shows that, for large path-loss exponents ($\alpha > 6$), every cut is approximately achievable, i.e., for every cut there exists an achievable unicast traffic matrix that has a sum-rate across the cut that is not much smaller than the cut capacity.

Lemma 8. *Under fast fading, for every $\alpha > 6$, there exists $b_4(n) \leq n^{o(1)}$ and $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}(n)$ such that for any n, $V(n) \in \mathcal{V}(n)$, and $S \subset V(n)$,*

$$C(S, S^c) \leq b_4(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w}. \tag{28}$$

Moreover, there exists a collection of channel gains $\mathcal{H}(n)$ such that

$$\mathbb{P}\big(\{h_{u,v}\}_{u,v \in V(n)} \in \mathcal{H}(n)\big) \geq 1 - o(1)$$

as $n \to \infty$, and such that for $\{h_{u,v}\}_{u,v} \in \mathcal{H}(n)$, (28) holds for slow fading as well.

Proof. By Lemma 7, for $V \in \mathcal{V}$

$$C(S, S^c) \leq K_1 \log^4(n) |\{v \in S^c : r_{S,v} < \log(n) + 1\}|, \tag{29}$$

where

$$r_{S,v} \triangleq \min_{u \in S} r_{u,v}.$$

Construct a unicast traffic matrix $\lambda^{\mathsf{UC}} \in \mathbb{R}_+^{n \times n}$ as

$$\lambda^{\mathsf{UC}}_{u,w} \triangleq \begin{cases} \rho(n) & \text{if } r_{u,w} < \log(n) + 1, \\ 0 & \text{else,} \end{cases}$$

for some function $\rho(n)$. We now argue that for $\rho(n) = \Theta(\log^{-2}(n))$ there exists $\tilde{b}(n) \geq n^{-o(1)}$ such that $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$. This follows from [19, Theorem 1] (see also Section IX.C there), once we show that for every $\ell \in \{1, \ldots, L(n)\}$ and $i \in \{1, \ldots, 4^\ell\}$ we have

$$\sum_{u \in V_{\ell,i}} \sum_{w \notin V_{\ell,i}} \lambda^{\mathsf{UC}}_{u,w} \leq (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2},$$

$$\sum_{u \notin V_{\ell,i}} \sum_{w \in V_{\ell,i}} \lambda^{\mathsf{UC}}_{u,w} \leq (4^{-\ell} n)^{2 - \min\{3,\alpha\}/2},$$

and for all $w \in V$

$$\sum_{u \neq w} \lambda^{\mathsf{UC}}_{u,w} \leq 1,$$

$$\sum_{u \neq w} \lambda^{\mathsf{UC}}_{w,u} \leq 1.$$

Since we assume that $V \in \mathcal{V}$, we have for all $w \in V$

$$\sum_{u \neq w} \lambda^{\mathsf{UC}}_{u,w} \leq K \log^2(n) \rho(n),$$
$$\sum_{u \neq w} \lambda^{\mathsf{UC}}_{w,u} \leq K \log^2(n) \rho(n),$$

for some constant K. By the locality of the traffic matrix λ^{UC}, it can be shown that this is sufficient for [19, Theorem 1] to apply with $\rho(n) = \frac{1}{K} \log^{-2}(n)$. Hence $\tilde{b}(n)\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ for fast fading, and the same holds for slow fading for some $\mathcal{H}$ with

$$\mathbb{P}\big(\{h_{u,v}\}_{u,v \in V} \in \mathcal{H}\big) \geq 1 - o(1)$$

as $n \to \infty$.

Combined with (29), this implies that

$$C(S, S^c) \leq \frac{K \log^6(n)}{\tilde{b}(n)} \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w},$$

proving the lemma. □

We are now ready for the proof of the outer bound on $\Lambda^{\mathsf{CA}}(n)$.

Proof of Theorem 6. Consider a cut $S \subset V$ in the wireless network. Assume we allow the nodes on each side of the cut to cooperate without any restriction — this can clearly only increase $\rho_{\lambda^{\mathsf{CA}}}$. The total amount of traffic that needs to be transmitted across the cut is then

$$\sum_{U \subset S} \sum_{w \notin S} \lambda^{\mathsf{CA}}_{U,w}.$$

The maximum achievable sum-rate (with the aforementioned node cooperation) is given by $C(S, S^c)$, the MIMO capacity between the nodes in S and in S^c. Therefore

$$\rho_{\lambda^{\mathsf{CA}}} \leq \min_{S \subset V} \frac{C(S, S^c)}{\sum_{U \subset S} \sum_{w \notin S} \lambda^{\mathsf{CA}}_{U,w}}. \tag{30}$$

We proceed by relating the cut S in the wireless network to a cut $\widetilde{S}$ in G. By Lemma 8, for $V \in \mathcal{V}$, there exists $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ such that for fast fading

$$C(S, S^c) \leq b_4(n) \sum_{u \in S} \sum_{w \notin S} \lambda^{\mathsf{UC}}_{u,w}, \tag{31}$$

and (31) holds also for slow fading if $\{h_{u,v}\}_{u,v} \in \mathcal{H}$ (with $\mathcal{H}$ defined as in Lemma 8). By [19, Theorem 1] (see again the discussion in Section IX.C there), for $\alpha > 5$ and $V \in \mathcal{V}$, there exists

K such that if $\lambda^{\mathsf{UC}} \in \Lambda^{\mathsf{UC}}$ then $K\log^{-6}(n)\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}}$, where G is the tree graph defined in Section III-A.

Now, consider any $\widetilde{S} \subset V_G$ such that $\widetilde{S} \cap V = S$. Note that $\widetilde{S}$ is a cut in G separating S from $V \setminus S$. Since $K\log^{-6}(n)\lambda^{\mathsf{UC}} \in \Lambda_G^{\mathsf{UC}}$, we thus have

$$\sum_{u\in S}\sum_{w\notin S} K\log^{-6}(n)\lambda^{\mathsf{UC}}_{u,w} \leq \sum_{\substack{(u,v)\in E_G:\\ u\in\widetilde{S},v\notin\widetilde{S}}} c_{u,v},$$

and by minimizing over the choice of $\widetilde{S}$ such that $\widetilde{S} \cap V = S$, we obtain

$$\sum_{u\in S}\sum_{w\notin S} K\log^{-6}(n)\lambda^{\mathsf{UC}}_{u,w} \leq \min_{\widetilde{S}:\widetilde{S}\cap V=S} \sum_{\substack{(u,v)\in E_G:\\ u\in\widetilde{S},v\notin\widetilde{S}}} c_{u,v}. \tag{32}$$

Combining (31) and (32) shows that

$$C(S,S^c) \leq \frac{b_4(n)}{K}\log^6(n) \min_{\widetilde{S}:\widetilde{S}\cap V=S} \sum_{\substack{(u,v)\in E_G:\\ u\in\widetilde{S},v\notin\widetilde{S}}} c_{u,v}.$$

Together with (30), and using Lemmas 4 and 8, this yields that with probability

$$\mathbb{P}(\{h_{u,v}\}_{u,v} \in \mathcal{H}, V \in \mathcal{V}) \geq 1 - o(1)$$

as $n \to \infty$, we have for any caching traffic matrix λ^{CA}

$$\begin{aligned}
\rho_{\lambda^{\mathsf{CA}}} &\leq \min_{S\subset V} \frac{C(S,S^c)}{\sum_{U\subset S}\sum_{w\notin S}\lambda^{\mathsf{CA}}_{U,w}} \\
&\leq b_2(n) \min_{S\subset V} \min_{\widetilde{S}\in V_G:\widetilde{S}\cap V=S} \frac{\sum_{\substack{(u,v)\in E_G:\\ u\in\widetilde{S},v\notin\widetilde{S}}} c_{u,v}}{\sum_{U\subset\widetilde{S}\cap V}\sum_{w\in V\setminus\widetilde{S}}\lambda^{\mathsf{CA}}_{U,w}} \\
&= b_2(n) \min_{\widetilde{S}\subset V_G} \frac{\sum_{\substack{(u,v)\in E_G:\\ u\in\widetilde{S},v\notin\widetilde{S}}} c_{u,v}}{\sum_{U\subset\widetilde{S}\cap V}\sum_{w\in V\setminus\widetilde{S}}\lambda^{\mathsf{CA}}_{U,w}} \\
&= b_2(n)\hat{\rho}_{\lambda^{\mathsf{CA}}},
\end{aligned}$$

with

$$b_2(n) \triangleq \frac{b_4(n)}{K}\log^6(n) \leq n^{o(1)}.$$

□

V. CONCLUSIONS

We analyzed the influence of caching on the performance of wireless networks. Our approach is information theoretic, yielding an inner bound on the caching capacity region for all values $\alpha > 2$ of path-loss exponent, and a matching (in the scaling sense) outer bound for $\alpha > 6$. Thus, in the high path-loss regime $\alpha > 6$, this provides a scaling characterization of the complete caching capacity region. Even though this region is $2^n \times n$-dimensional (i.e., exponential in the number of nodes n in the wireless network), we present an algorithm that checks approximate feasibility of a particular caching traffic matrix efficiently (in polynomial time in the description length of the caching traffic matrix). Achievability is proved using a three-layer communication architecture achieving the entire caching capacity region in the scaling sense for $\alpha > 6$. The three layers deal with optimal selection of caches, choice of amount of necessary cooperation, noise and interference, respectively. The matching (in the scaling sense) converse proves that addressing these questions separately is without loss of order-optimality in the high path-loss regime.

REFERENCES

[1] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, March 2000.

[2] P. Nuggehalli and C.-F. Chiasserini V. Srinivasan. Energy-efficient caching strategies in ad hoc wireless networks. In *ACM international symposium on Mobile Ad Hoc Networking & Computing*, pages 25–34, 2003.

[3] S. Bhattacharya, H. Kim, S. Prabh, and T. Abdelzaher. Energy-conserving data placement and asynchronous multicast in wireless sensor networks. In *International Conference On Mobile Systems, Applications And Services*, pages 173–185, 2003.

[4] S. Jin and L. Wang. Content and service replication strategies in multi-hop wireless mesh networks. In *International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems*, pages 79–86, 2005.

[5] B.-J. Ko and D. Rubenstein. Distributed self-stabilizing placement of replicated resources in emerging networks. *IEEE Transactions on Networking*, 13(3):476–487, June 2005.

[6] L. Yin and G. Cao. Supporting cooperative caching in ad hoc networks. *IEEE Transactions on Mobile Computing*, 5(1):77–89, January 2005.

[7] T. Cover, A. E. Gamal, and M. Salehi. Multiple access channels with arbitrarily correlated sources. *IEEE Transactions on Information Theory*, 26(6):648– 657, November 1980.

[8] T. S. Han. Slepian-Wolf-Cover theorem for network of channels. *Inform. and Control*, 47(1):67–83, January 1980.

[9] J. Barros and S. D. Servetto. Network information flow with correlated sources. *IEEE Transactions on Information Theory*, 52(1):155–170, January 2006.

[10] L.-L. Xie and P. R. Kumar. A network information theory for wireless communication: Scaling laws and optimal operation. *IEEE Transactions on Information Theory*, 50(5):748–767, May 2004.

[11] A. Jovičić, P. Viswanath, and S. R. Kulkarni. Upper bounds to transport capacity of wireless networks. *IEEE Transactions on Information Theory*, 50(11):2555– 2565, November 2004.

[12] O. Lévêque and İ. E. Telatar. Information-theoretic upper bounds on the capacity of large extended ad hoc wireless networks. *IEEE Transactions on Information Theory*, 51(3):858–865, March 2005.

[13] F. Xue, L.-L. Xie, and P. R. Kumar. The transport capacity of wireless networks over fading channels. *IEEE Transactions on Information Theory*, 51(3):834–847, March 2005.

[14] L.-L. Xie and P. R. Kumar. On the path-loss attenuation regime for positive cost and linear scaling of transport capacity in wireless networks. *IEEE Transactions on Information Theory*, 52(6):2313–2328, June 2006.

[15] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran. Closing the gap in the capacity of wireless networks via percolation theory. *IEEE Transactions on Information Theory*, 53(3):1009–1018, March 2007.

[16] P. Gupta and P. R. Kumar. Towards an information theory of large networks: An achievable rate region. *IEEE Transactions on Information Theory*, 49(8):1877–1894, August 2003.

[17] A. Özgür, O. Lévêque, and D. N. C. Tse. Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *IEEE Transactions on Information Theory*, 53(10):3549–3572, October 2007.

[18] U. Niesen, P. Gupta, and D. Shah. On capacity scaling in arbitrary wireless networks. *submitted to IEEE Transactions on Information Theory*, November 2007. Available online at http://arxiv.org/abs/0711.2745.

[19] U. Niesen, P. Gupta, and D. Shah. The balanced unicast and multicast capacity regions of large wireless networks. *submitted to IEEE Transactions on Information Theory*, July 2008. Available online at http://arxiv.org/abs/0809.1344.

[20] M. Franceschetti, M. D. Migliore, and P. Minero. The capacity of wireless networks; information-theoretic and physical limits. 55(8):3413–3424, August 2009.

[21] M. Franceschetti, M. D. Migliore, and P. Minero. The degrees of freedom of wireless networks: Information theoretic and physical limits. In *Allerton Conference on Communication, Control, and Computing*, September 2008.

[22] R. Srikant. *The Mathematics of Internet Congestion Control*. Birkhäuser, 2003.

[23] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, June 1995.

[24] A. Agarwal, N. Alon, and M. Charikar. Improved approximation for directed cut problems. In *ACM Symposium on Theory of Computing*, pages 671–680, 2007.

[25] N. Kahale. On reducing the cut ratio to the multicut problem. *Unpublished Manuscript*, 1993.

[26] N. Garg, V. V. Vazirani, and M. Yannakakis. *Primal-dual approximation algorithms for integral flow and multicut in trees, with applications to matching and set cover*, pages 64–75. Lecture Notes in Computer Science. Springer, 1993.